Principal Component Analysis

Dataset df.pca

This dataset contains 36 individuals and 19 variables, 1 qualitative variable is considered as illustrative.

### 1. Study of the outliers

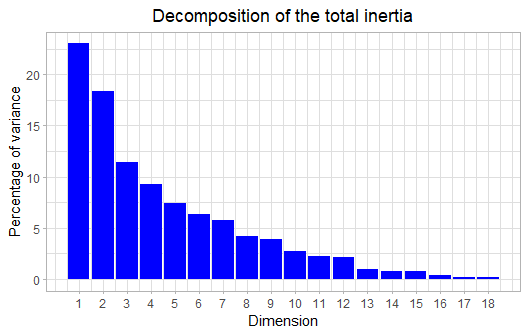
The analysis of the graphs does not detect any outlier.

### 2. Inertia distribution

The inertia of the first dimensions shows if there are strong relationships between variables and suggests the number of dimensions that should be studied.

The first two dimensions of analyse express **41.38%** of the total dataset inertia ; that means that 41.38% of the individuals (or variables) cloud total variability is explained by the plane. This is an intermediate percentage and the first plane represents a part of the data variability. This value is greater than the reference value that equals **28.51%**, the variability explained by this plane is thus significant (the reference value is the 0.95-quantile of the inertia percentages distribution obtained by simulating 1426 data tables of equivalent size on the basis of a normal distribution).

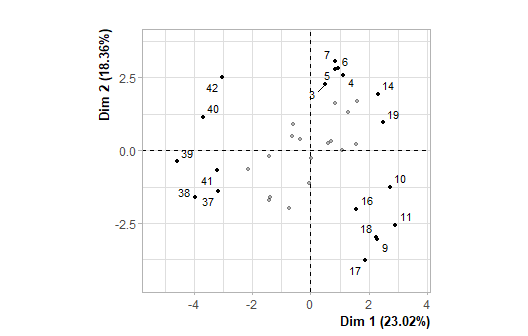
From these observations, it may be interesting to consider the next dimensions which also express a high percentage of the total inertia.



**Figure 2 - Decomposition of the total inertia**

An estimation of the right number of axis to interpret suggests to restrict the analysis to the description of the first 3 axis. These axis present an amount of inertia greater than those obtained by the 0.95-quantile of random distributions (52.78% against 39.06%). This observation suggests that only these axis are carrying a real information. As a consequence, the description will stand to these axis.

### 3. Description of the plane 1:2

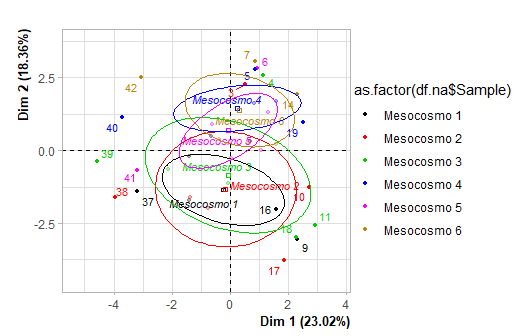


**Figure 3.1 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

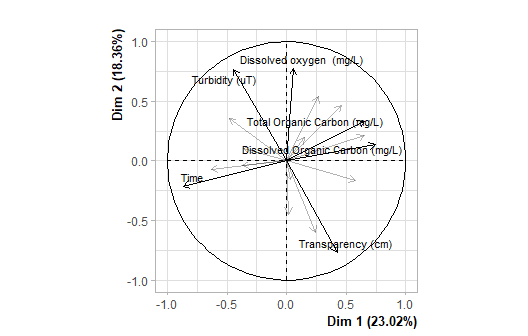
The Wilks test p-value indicates which variable factors are the best separated on the plane (i.e. which one explain the best the distance between individuals).

## as.factor(df.na$Sample)   
## 0.04475331

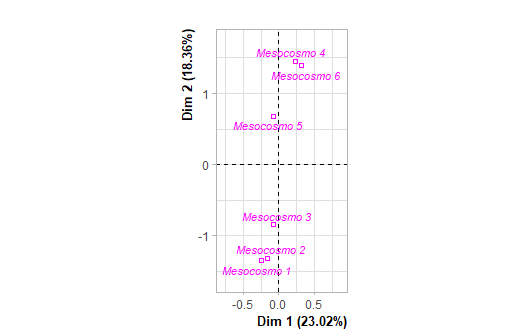
There only is one possible qualitative variable to illustrate the distance between individuals : *as.factor(df.na$Sample)*.



**Figure 3.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.* *The individuals are coloured after their category for the variable* as.factor(df.na$Sample).



**Figure 3.3 - Variables factor map (PCA)** *The labeled variables are those the best shown on the plane.*



**Figure 3.4 - Qualitative factor map (PCA)** *The labeled factors are those the best shown on the plane.*

The **dimension 1** opposes individuals such as *11*, *14*, *10*, *7*, *5*, *9*, *6*, *17*, *18* and *19* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *40*, *37*, *38*, *39*, *42* and *41* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group in which the individuals *14*, *7*, *5*, *6* and *19* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for the variables *Fluoride.(mg/L)*, *Dissolved.oxygen..(mg/L)*, *Total.Organic.Carbon.(mg/L)*, *Dissolved.Organic.Carbon.(mg/L)* and *Organic.matter.(254.Abs)* (variables are sorted from the strongest).
* low values for the variables *Time*, *Percentual* and *Chloride.(mg/L)* (variables are sorted from the weakest).

The group in which the individuals *11*, *10*, *9*, *17* and *18* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for the variables *Transparency.(cm)*, *Percentual*, *Orthophosphate.(mg/L)* and *Conductivity.(µS/cm)* (variables are sorted from the strongest).
* low values for the variables *Turbidity.(uT)*, *Nitrate.(mg/L)*, *Dissolved.oxygen..(mg/L)* and *Time* (variables are sorted from the weakest).

The group in which the individuals *40*, *37*, *38*, *39*, *42* and *41* stand (characterized by a negative coordinate on the axis) is sharing :

* high values for the variables *Time*, *Chloride.(mg/L)* and *Nitrate.(mg/L)* (variables are sorted from the strongest).
* low values for the variables *Dissolved.Organic.Carbon.(mg/L)*, *Conductivity.(µS/cm)*, *Total.Organic.Carbon.(mg/L)*, *Orthophosphate.(mg/L)*, *Fluoride.(mg/L)* and *Transparency.(cm)* (variables are sorted from the weakest).

The **dimension 2** opposes individuals such as *14*, *7*, *5*, *6* and *19* (to the top of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *11*, *10*, *9*, *17* and *18* (to the bottom of the graph, characterized by a strongly negative coordinate on the axis).

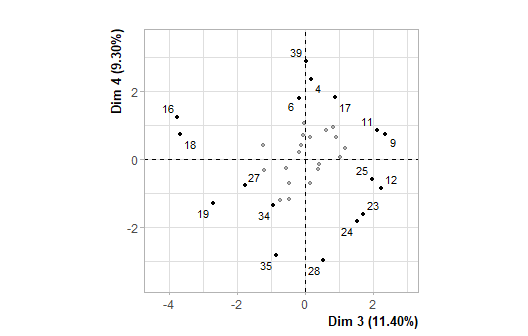
The group in which the individuals *14*, *7*, *5*, *6* and *19* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for the variables *Fluoride.(mg/L)*, *Dissolved.oxygen..(mg/L)*, *Total.Organic.Carbon.(mg/L)*, *Dissolved.Organic.Carbon.(mg/L)* and *Organic.matter.(254.Abs)* (variables are sorted from the strongest).
* low values for the variables *Time*, *Percentual* and *Chloride.(mg/L)* (variables are sorted from the weakest).

The group in which the individuals *11*, *10*, *9*, *17* and *18* stand (characterized by a negative coordinate on the axis) is sharing :

* high values for the variables *Transparency.(cm)*, *Percentual*, *Orthophosphate.(mg/L)* and *Conductivity.(µS/cm)* (variables are sorted from the strongest).
* low values for the variables *Turbidity.(uT)*, *Nitrate.(mg/L)*, *Dissolved.oxygen..(mg/L)* and *Time* (variables are sorted from the weakest).

### 4. Description of the dimension 3

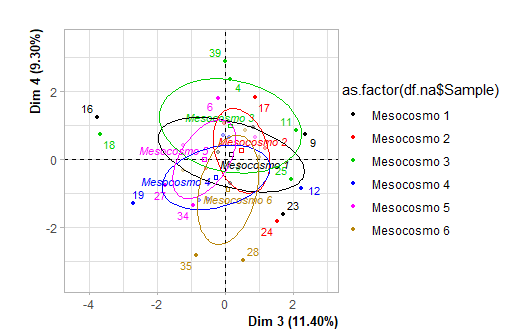


**Figure 4.1 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.*

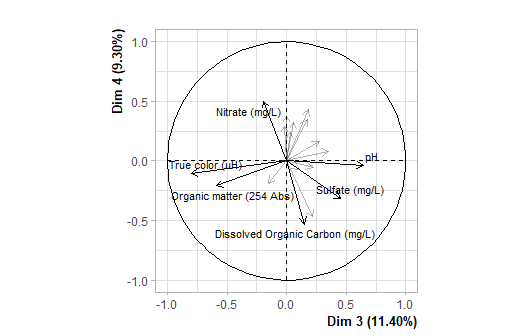
The Wilks test p-value indicates which variable factors are the best separated on the plane (i.e. which one explain the best the distance between individuals).

## as.factor(df.na$Sample)   
## 0.4977486

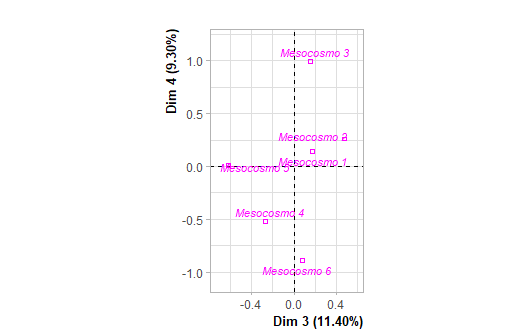
There only is one possible qualitative variable to illustrate the distance between individuals : *as.factor(df.na$Sample)*.



**Figure 4.2 - Individuals factor map (PCA)** *The labeled individuals are those with the higher contribution to the plane construction.* *The individuals are coloured after their category for the variable* as.factor(df.na$Sample).



**Figure 4.3 - Variables factor map (PCA)** *The labeled variables are those the best shown on the plane.*



**Figure 4.4 - Qualitative factor map (PCA)** *The labeled factors are those the best shown on the plane.*

The **dimension 3** opposes individuals such as *24*, *23*, *12*, *9* and *11* (to the right of the graph, characterized by a strongly positive coordinate on the axis) to individuals such as *16*, *19*, *18*, *27* and *34* (to the left of the graph, characterized by a strongly negative coordinate on the axis).

The group in which the individuals *9* and *11* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for the variables *Transparency.(cm)*, *pH* and *Orthophosphate.(mg/L)* (variables are sorted from the strongest).

The group in which the individuals *24*, *23* and *12* stand (characterized by a positive coordinate on the axis) is sharing :

* high values for the variables *Temperature.(°C)*, *Nitrite.(mg/L)* and *Chloride.(mg/L)* (variables are sorted from the strongest).

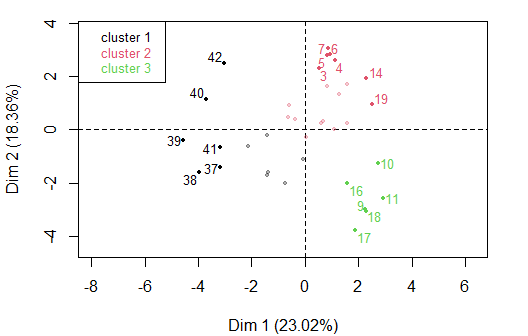
The group in which the individuals *16* and *18* stand (characterized by a negative coordinate on the axis) is sharing :

* high values for the variable *True.color.(uH)*.
* low values for the variables *Turbidity.(uT)* and *pH* (variables are sorted from the weakest).

The group in which the individuals *19*, *27* and *34* stand (characterized by a negative coordinate on the axis) is sharing :

* low values for the variables *Percentual* and *Chloride.(mg/L)* (variables are sorted from the weakest).

### 5. Classification



**Figure 5 - Ascending Hierarchical Classification of the individuals.** *The classification made on individuals reveals 3 clusters.*

The **cluster 1** is made of individuals such as *37*, *38*, *39*, *40*, *41* and *42*. This group is characterized by :

* high values for the variables *Time*, *Chloride.(mg/L)*, *Nitrite.(mg/L)* and *Sulfate.(mg/L)* (variables are sorted from the strongest).
* low values for the variables *Conductivity.(µS/cm)*, *Total.Organic.Carbon.(mg/L)*, *Dissolved.Organic.Carbon.(mg/L)*, *Organic.matter.(254.Abs)*, *Fluoride.(mg/L)* and *Orthophosphate.(mg/L)* (variables are sorted from the weakest).

The **cluster 2** is made of individuals such as *3*, *4*, *5*, *6*, *7*, *14* and *19*. This group is characterized by :

* high values for the variables *Total.Organic.Carbon.(mg/L)*, *Organic.matter.(254.Abs)*, *Dissolved.oxygen..(mg/L)*, *Dissolved.Organic.Carbon.(mg/L)* and *Fluoride.(mg/L)* (variables are sorted from the strongest).
* low values for the variables *Chloride.(mg/L)*, *Percentual*, *Transparency.(cm)* and *Time* (variables are sorted from the weakest).

The **cluster 3** is made of individuals such as *9*, *10*, *11*, *16*, *17* and *18*. This group is characterized by :

* high values for the variables *Transparency.(cm)*, *Percentual*, *Orthophosphate.(mg/L)* and *Conductivity.(µS/cm)* (variables are sorted from the strongest).
* low values for the variables *Turbidity.(uT)*, *Dissolved.oxygen..(mg/L)* and *Nitrate.(mg/L)* (variables are sorted from the weakest).

## Annexes